22.3 Fine-Grain Redundant Logic Using Defect-Prediction Flip-Flops

Toru Nakura, Koichi Nose, Masayuki Mizuno

NEC, Kanagawa, Japan

As CMOS process technology advances, an initial failure compensation technique for yield enhancement, and an in-field failure prevention technique against increasing in-field failure rates are becoming crucial, especially in such highly-dependable chip market as automotive. This paper introduces fine-grain redundant logic (FGR) for switching only a defective portion of a main circuit having killer/latent defects to its redundant sub-circuit block, and defect prediction flip-flops (DPFFs) for the autonomous prediction and prevention of in-field failures due to latent defects, as shown in Fig. 22.3.1. Here, killer defects refer to those apparent at fabrication process, while latent defects refer to those apparent only in actual use. While redundant circuits to compensate defect-related multi-failures have been widely used for memory, we introduce redundancy into logic circuits as well. The DPFFs are connected via the scan chain through which defect points are traced. The scan chain is connected to an external nonvolatile memory so as to reproduce main or sub switching states on the basis of screening tests and in-field failure prevention results. Our estimates show that the application of this scheme enhances a production yield of, for example, 70% to 91% even with considering the area increase due to the redundancy, and it prevents over 80% of in-field failures caused by one or two latent defects on a chip.

Chips with latent defects pass a screening test only to fail in early operations. Examples of latent defects are partial insufficiency, cracked vias, extra-metal, etc., which can develop into the opening or shorting of connections. Most latent defect-related failures are predictable since they gradually appear as path delay increases in actual use, as shown in Fig. 22.3.2. Here we introduce *prediction time* as the short period of time just before the boundary between correct and failed operations, and a warning is issued if the path delay falls within this period. The allowed maximum path delay for correct operations stays the same while the allowed maximum path delay for warning-free operations, which is the timing constraint in our designs, decreases by the prediction time. This prediction is realized by a DPFF that has i) a system flip-flop to maintain normal operations, ii) a delay line to determining the prediction time, iii) a warning generation XOR, and iv) a warning-hold/scannable latch, as shown in Fig. 22.3.2. Also, the DPFF has dual input/output and their control nodes for duplicate logic.

Figure 22.3.3 shows our FGR using DPFF scheme. All the logic elements are duplicated, with main- and sub-circuit blocks. Normally, sub outputs of DPFFs are off and DPFFs use data inputs from the main-circuit blocks, which means that only main-circuit blocks operate. Referring to Fig. 22.3.3, the block previous to DPFF(a) has a latent defect that creates a path delay increase causing the following steps: 1) DPFF(a) detects the path delay increase and generates a warning, 2) the warning is distributed to DPFF(b), DPFF(c) and DPFF(d) whose outputs determine DPFF(a) data input, 3) the DPFFs receiving the warning activate their respective sub output, and the following redundant sub-blocks, which have no defect, start to work. DPFF(a) begins to use input data from the sub-block, since output port W and main/sub input selection node (M/S in Fig. 22.3.2) are connected inside each DPFF here so as to change the input data port once a warning is issued. These procedures cause no system disturbance since the switching from main to sub is conducted while the main block still maintains correct operation. The warning distribution is restricted to only DPFFs whose outputs determine the input of the DPFF which issued the warning, so that the blocks to be switched are partitioned into fine grains. In all other areas, only main-blocks operate as usual so that the sub-blocks consume no active power. While the DPFF is close to Razor [1] or perturbation tolerant/detecting circuits [2], both of which use time-redundant techniques for error detection and correction, the DPFF predicts and prevents failures before defects actually cause errors, thereby our circuit maintains correct operations using both time and circuit redundancy.

As shown in Fig. 22.3.4, DPFFs with MUX-scan are 3.15× larger than normal DFFs with MUX-scan, and the combination-logic area for a FGR design is more than twice as large due to its redundancy and warning distribution OR gates. If the area ratio of combination-logic/DFF is 6:4, the area becomes about 2.5× larger. Considering, however, the ITRS prediction that the area ratio on an SoC for logic/memory will reduce, as the process technology advances to 12/88 at a 45nm node, for example, the overall area penalty is only 18%, as shown in Fig. 22.3.4. A chip yield, with respect to the logic area of, for example, 70% would increase to 91% under the assumption that a) FGR designs have more than 1000 grains within a logic area being increased by 2.5× and b) 10% of defects are related to DPFFs which do not have redundancy. This number 10% is arrived at assuming that DPFFs occupy 50% of the logic area and that the portion of defects occurring under the M1 layer in a multi-metal layer process is 20% (0.5×0.2=0.1). Furthermore, more than 80% in-field failures due to two latent defects are prevented, while a triplicate scheme with a winner-take-all circuit would prevent only 33%, and single and duplicate schemes would prevent none, as shown in Fig. 22.3.4. These yield enhancement and in-field failure reductions prevail over the disadvantage of the chip area increase.

To evaluate our scheme, we first design and fabricate failure prediction flip-flops with a pseudo-defect circuit, using 90nm standard CMOS. A chip micrograph is shown in Fig. 22.3.7. The pseudo-defect circuit in Fig. 22.3.5 imitates the metal insufficiency shown in Fig. 22.3.2. The waveforms of QOUT and WOUT are observed with a sampling oscilloscope at 330MHz operation with Vstress bias conditions indicated as (i)-(iv), as shown in Fig. 22.3.5. We see that as Vstress increases, (i) correct operations with no warning, (ii),(iii) warning is issued and still maintains correct operations until (iv) QOUT outputs a wrong signal. Here, the prediction time is 95ps.

We also fabricate a 16b digital ΔΣ modulator with our FGR/DPFF, as shown in Fig. 22.3.7. The pseudo-defect circuits are inserted at two points, and we control if the pseudo-defects are ON or OFF as well as FGR function is ON or OFF. Figure 22.3.6 shows measured waveforms of CLK and ΔΣ output waveforms of defect:OFF as a reference, defect:ON/FGR:ON and defect:ON/FGR:OFF cases. The pseudo defects appear at 150ns. In the operation mode, the FGR:ON case maintains correct operation even after the defects appear while the FGR:OFF case outputs the wrong signal. The scan output shows two pulses corresponding to the two inserted pseudo-defect circuits. We design a normal single/DFF version as well. As shown in the table in Fig. 22.3.6, in a normal single/DFF scheme, the ratio of the combinational-logic area and DFF area of the ΔΣ modulator is 0.61:0.39. In the FGR/DPFF scheme, combinational-logic area increases by 2.2×, and flip-flop area by 3.15×, with the increase in total area becoming 2.57×. The table also shows the combination-logic/DFF area ratio on three commercial chips, and for the FGR/DPFF scheme as estimated using the multiples 2.2× and 3.15×, with the number of grids (reflecting the chip size) derived from respective logic synthesis reports. Assuming the ITRS 45nm logic/memory ratio, the SoC area penalty would be only less than 20%.

References:

[1] S. Das, D. Roberts, S. Lee, et al., "A Self-Tuning DVS Processor Using Delay-Error Detection and Correction," *IEEE J. Solid-State Circuits,* vol. 41, no. 4, pp. 792-804, Apr., 2006.

[2] M. Nicolaidis, "Time Redundancy Based Soft-Error Tolerance to Rescue Nanometer Technologies," *IEEE VLSI Test Symposium*, pp. 86-94, Apr., 1999.

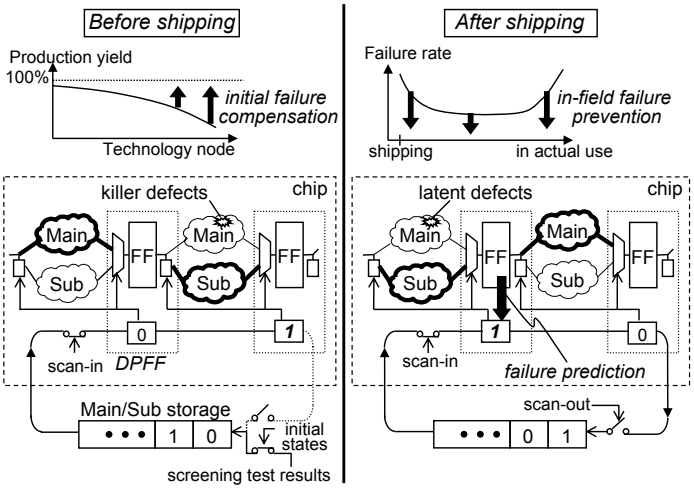


Figure 22.3.1: Yield enhancement and in-field failure rate reduction.

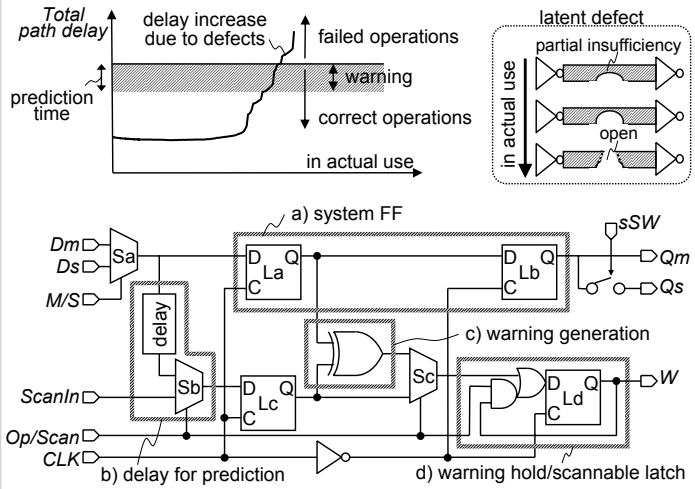


Figure 22.3.2: Path delay increase and DPFF.

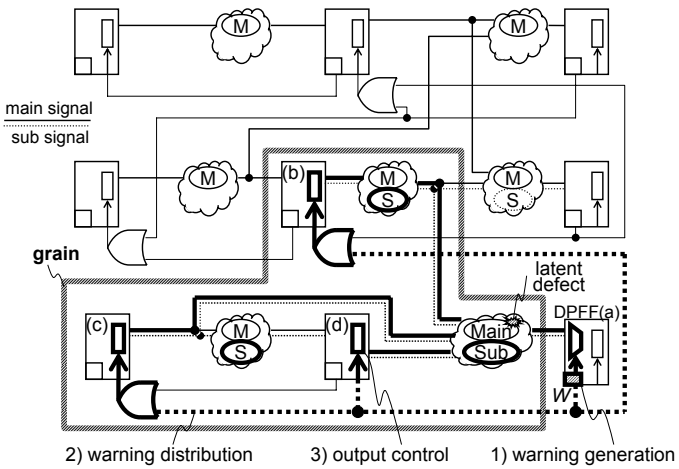


Figure 22.3.3: Fine-grain redundancy using DPFF.

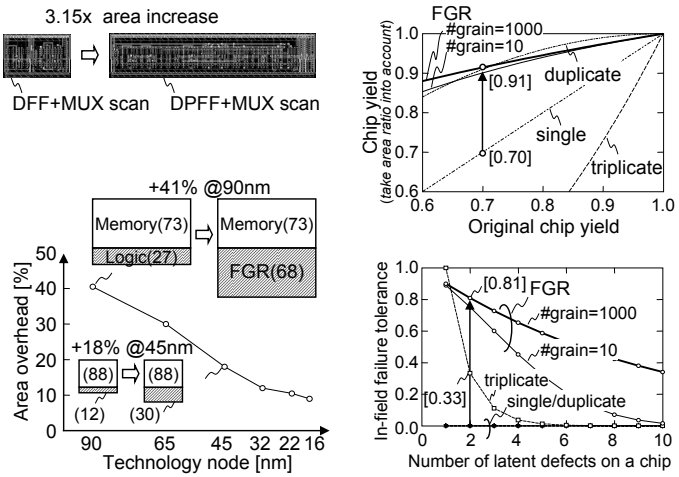


Figure 22.3.4: Area penalty, yield and in-field failure tolerance.

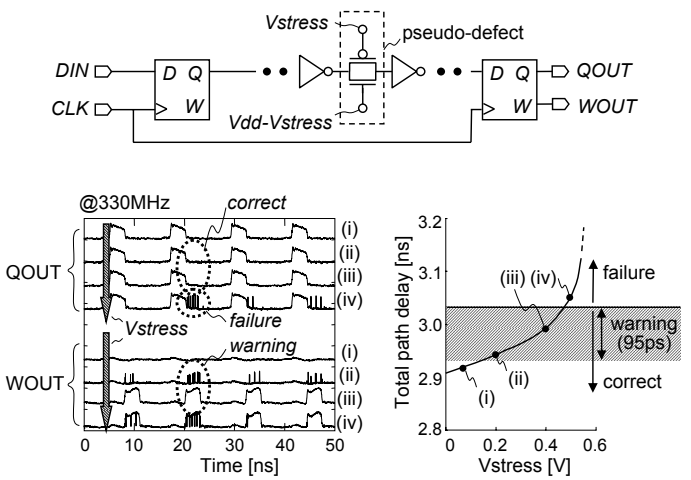
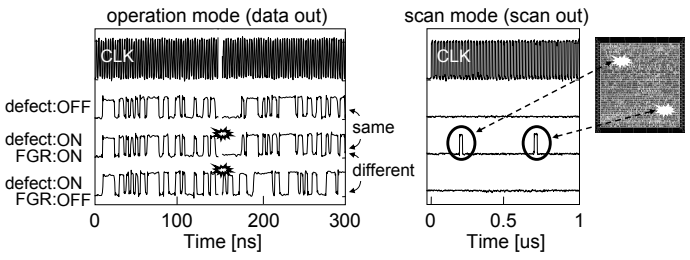


Figure 22.3.5: Test circuit, and measured waveforms.

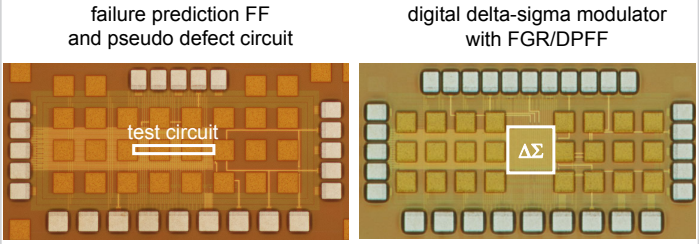


Area Overhead

	chipA (ΔΣ)	chipB	chipC	chipD
Number of grids	3.3K	25.6M	20.7M	2.2M
Area ratio of Comb.-Logic : DFF	0.61 : 0.39 (2.2x ↓ + ↓ 3.15x)	0.63:0.37	0.69:0.31	0.56:0.44
Area increase in FGR	2.57	2.55	2.49	2.62
Chip total area overhead (Logic:Memory=12:88@45nm)	18.8%	18.6%	17.5%	19.4%

Figure 22.3.6: Measured waveforms and area overhead.

Continued on Page 611



Technology	90nm 1P6M CMOS
Chip size	1.65mm x 0.80mm
Supply Voltage	1.0V

Figure 22.3.7: Chip micrographs.

